**DECLARATION**

We Divyanka Latwal, Karan Dawar, Shivani Singh and Surabhi Nanda, students of B.Tech (CSE) hereby declare that the Major project entitled *Budget 2018 Analysis using Tweets* which is submitted to Department of Computer Science HMR Institute of Technology & Management, Hamidpur Delhi, affiliated to Guru Gobind Singh Indraprastha University, Dwarka(New Delhi) in partial fulfillment of requirement for the award of the degree of Bachelor of Technology in CSE, has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition. The list of member(s) involved in the project is listed below:-

S. No. Student Name Enrollment Number Student Signature

1. Divyanka Latwal 12813302714

2. Karan Dawar 30413302714

3. Shivani Singh 11813302714

4. Surabhi Nanda 60413302714

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

New Delhi **GUIDE:**

Date: 03/10/2017 Ms. CHITRA NASA

Assistant Professor

Department of CSE

Dr. Ravinder Kumar

HOD, Department of CSE

HMRITM Hamidpur, New Delhi

**ACKNOWLEDGEMENT**

“A Teacher is a compass that activates the magnets of curiosity, wisdom, and knowledge in hearts of people. The completion of any inter-disciplinary Training project depends upon cooperation, coordination and combined efforts of several sources of knowledge. Gratitude is not a thing of expression; it is more matter of feeling. There is always a sense of gratitude which one express towards others for their help and supervision in achieving the goals. This formal piece of acknowledgment is an attempt to express the feeling of Gratitude towards people who help me in successfully completing my training.

I would like to express my deep gratitude to Ms. Chitra Nasa, my coordinator for her Constant co-operation. She was always there with her competent guidance and valuable suggestion throughout the pursuance of the project.

I am also thankful to all the respondents whose cooperation & support has Helped me a lot in collecting necessary information.

**ABSTRACT**

With the advancement of web technology and its growth, there is a huge volume of data present in the web for internet users and a lot of data is generated too. Internet has become a platform for online learning, exchanging ideas and sharing opinions. Social networking sites like Twitter, Facebook, Google+ are rapidly gaining popularity as they allow people to share and express their views about topics, have discussion with different communities, or post messages across the world. Twitter, one of the largest social media site receives tweets in millions every day. This huge amount of raw data can be used for industrial or business purpose by organizing according to our requirement and processing. This project provides a way of Budget 2018 analysis using Tweets. Analyzing this vast quantity of unstructured data presents challenges for software and hardware.

This project addresses the problem of sentiment analysis in twitter; that is classifying tweets according to the sentiment expressed in them: positive, negative or neutral. The aim of this project is to develop a functional classifier for accurate and automatic sentiment classification of an unknown tweet stream. To build a model that can analyses the response of people on particular topic.

**CONTENT**

**Declaration………….……………………………………………………………….………..i**

**Acknowledgement……....……………………………………………………………………ii**

**Abstract…………………………..…………………………………………………………..iii**

**Contents…………………………………………………………………………………...…iv**

**List of Figures………………………………………………………………………..………v**

**Chapter 1:Introduction ………………………..…………………………………………….1**

* 1. Micro Blogging …………………………………………………………...…......1
  2. How does twitter work ...…………………………………………..………...…..2
  3. Who uses twitter and why ...……………………………………...…………...…4

**Chapter 2: Data Collection** **…………………………………………………………….........6**

2.1 Why Twitter ………………………………………………………………..........6

2.2 Types of API …………………………………………………………………….6

2.2.1 REST APIs ………………………………………………………………...6

2.2.2 Streaming APIs ……………………………………………………………6

2.3 Outh Authentication ……….……………………………………………………7

**Chapter 3: Data Processing………………………………………………………………...10**

3.1 Pre-Processing Textual Data ……..…………………………..…………...……10

3.2 Transformation ...…...………………………………………………..…………10

3.2.1. HTML Cleanup ...…...……………………………… ...…...……………10

3.2.2. Expanding Abbreviation ...…...……………...…………………………...11

3.2.3. White Space And Stop words Removal ...…..……………………………11

3.2.4 Handling Negation ….……………………...……………………………..13

**Chapter 4: Analysis………………………………………………………………................15**

4.1 Word Cloud ………………..……..…………………………..…………...……15

4.2 Hierarchical Clustering ...…...………………………………………..…………16

4.3 Topic Modeling …...……………………………………………………………20

4.3.1 Latent Dirichlet allocation ….....……………………….…………………21

4.4 Sentiment Analysis…………………………………………...………...……….22

4.4.1. Differences in Definition ………………...………...…………………….22

**Conclusion**………………….………………………………………………………………..36

**References**……………………………………………………………………………….......37

**LIST OF FIGURES**

**Name of Figure Page No.**

Fig. 2.1 : Outh WorkFlow…………………………………………………………….. 9

Fig 3.1 Removing undesirable texts …………………………………………………. 14

Fig 4.1 Agnes versus Diana …………………………………………………………... 17

Fig 4.2 Different approaches of Dendrogram ………………………………………... 19

Fig 4.3 Topic Modeling Analysis ……………………………………………………. 20

Fig 4.4 Dyadic implications …………………………………………………………. 28

Fig A1 Home Screen …………………………………………………………………. 29

Fig A2 Word Cloud …………………………………………………………..………. 29

Fig A3 Sentiment Analysis…………………………………………………...………… 30

Fig A4 Hierarchical Clustering ……………………………………………...………… 30

Fig A5 Sentiments Trends …………………………………………………...………… 31

Fig A6 Sentiment and Emotion Analysis …………………………………...………… 31